Tree Based Method for Censored Data

Anaranya Basu

Department Of Statistics

September 18, 2023

Supervised By
Prof. Biswabrata Pradhan
Indian Statistical Institute, Kolkata

Declaration

I hereby declare that the project entitled **Tree Based Method for Censored Data**, an Internship report, submitted in partial fulfillment for the degree of Master of Science (M.Sc) in Statistics, St.Xaviers University, completed under the supervision of Prof. Biswabrata Pradhan, at Indian Statistical Institute, Kolkata, is an authentic work. I also affirm that no part of my research project uses unacknowledged materials and resources.

Anaranya Basu
M.Sc Statistics
St. Xavier's University

Acknowledgement

Writing a project report takes a significant amount of time to be completed. In this period I have been enriched in numerous ways by interaction with many people, and at this point I would like to thank them all. But a specific cohort among them deserves my heartfelt mention.

First of all I would like to express my humble gratitude towards the person, perhaps a stalwart in his field, who has proposed the central idea of this project, that is my supervisor
Prof. Biswabrata Pradhan of Indian Statistical Institute, whose encouragement, guidance and support from the initial to final level enabled me to develop an understanding of the project.

I would like to express my humble gratitude towards the person, none other than **Prof.Dr.Manisha Pal** (HOD, Department of Statistics, St.Xavier's University). The idea of engraving concepts and creating a strong base is what she is teaching me throughout my M.Sc journey at the Department. And I am also thankful to my professor Dr.Priyanka Talukdar who taught me various complex concepts in a simplistic manner.

Abstract

A tree-based method for censored survival data is discussed, based on maximizing the difference in survival between groups of patients by the Log-Rank statistic value based criterion. The method includes a pruning algorithm with optimal properties analogous to the classification and regression tree (RPART) pruning algorithm. After that conventional tree based method is discussed with optimal pruning strategy. Then Concordance index measure is discussed and as a better accuracy measure Brier Score method is considered. And three example is given to show the utility of the algorithm for developing prognostic classifications for censored observation raised from Bio-medical field.

KEY WORDS : Censored Data; Classification and regression tree; Regression trees; Concordance Index; Brier score; Survival analysis .

This page intentionally left blank

Contents

1 Introduction

Survival analysis is a set of statistical models and methods used for estimating time until the occurance of an event (or the probability that an event has not occured). These methods are widely used in demography, e.g. for estimating lifespan or age at the first childbirth, in healthcare, e.g. for estimating duration of staying in a hospital or survival time after the diagnosis of a disease, in engineering (for reliability analysis), in insurance, economics, and social sciences. Statistical methods need data, but complete data may not be available, i.e. the exact time of the event may be unknown for certain reasons (the event did not occur before the end of the study or it is unknown whether it occured). In this case, lifetimes are called censored. The data are censored from below (left censored) when below a given value the exact values of observations is unknown. Right censored data (censored from above) does not have exact observations above a given value. Further in this paper, right censoring is considered.

The problems studied with the help of survival analysis are formulated in terms of survival function (that is complementary distribution function)

$$S(t) = P(T > t)$$

where t is observation time and T is random variable standing for event time. The distribution of T may also be characterized with so called hazard function,

$$h(t) = -\frac{\partial}{\partial t} \log s(t)$$

There are several ways for estimating the survival function. A parametric model assumes a distribution function, and its parameters are estimated based on the available data. Also we may find empirical distribution function and then use its complement as the survival function. Non parametric methods called the Kaplan-Meier estimator (Kaplan and Meier, 1958) and Nelson-Alan (Nelson, 1972) estimator are more powerful. The Kaplan-Meier estimator has the form,

$$S(t) = \prod_{i:t_i<t}(1 - \frac{d_i}{n_i})$$

where t_i is time of the event, d_i is number of events that occurred at time t_i, and n_i number of events after time t_i (or unknown at

t_i). Nelson-Aalen esimator applies the same idea to the cumulative hazard function $H(t) = \int_0^t h(s)ds$ and then transforms it to the estimation of the survival function.

Tree based approach is a well known supervised method which actually classifies data and partition a data set into smaller groups and then fit a simple model (constant) for each subgroup. In case of censored data where the data comes in a incomplete format tree based method is helpful sometimes to know the survival probability with some important risk factors. Regression trees are highly interperetable, which can be important for understanding the factors that influence survival or time-to-event outcomes in censored data. You can easily visualize and interpret the decision tree structure to gain insights into the relationships between predictors and survival.

2 Notation And Data Set up

We assume that data include failure time measurements and additional measurements (covariates) that may be associated with failure time. An observation will be distributed as the vector (T,δ,X), where T is the time under observation, δ is an indicator of failure, and X is a vector of M covariates. Let U denote the true survival time having cumulative distribution function F, and let V be the true censoring time with cumulative distribution function C. Then assume δ=$I_{U \leq V}$, where $I_{.}$ is the indicator function of the set . and define the observed time T = min(U, V). Assume also that U and V are independent given X, for identifiability reasons. The learning sample consists of the set of iid vectors :

$$(T_i, \delta_i, X_i) : i = 1, 2, \ldots, N$$

This whole process is done in R by the **Surv()** function.

- Surv(time, status)- right censored data
- Surv(time, endpoint==“death”) - right censored data, where the status variable is a character or factor.
- Surv(t1, t2, status) - counting process data
- Surv(t1, ind, type= “left”) - left censoring

3 Theoretical Part and Splitting for censored Data

Ciampi et al. (Ciampi et al., 1986) suggested to use the logrank statistic (Lee, 2021) for comparison of the two groups of observation in the children nodes. The more is the value of the statistic the more the hazard functions of the groups differ. The splitting is chosen for the largest statistic value. Leblanc M. and Crowley J. (LeBlanc and Crowley, 1993) introduced a tree algorithm based on logrank statistic in combination with the cost-complexity pruning algorithm.

The algorithm splits the data into groups with differing survival. Because the logrank test has been used extensively in the analysis

of censored survival data, it is a logical choice for measuring dissimilarity in survival between two groups. Let the number of patients in the two groups be n_1, and n_2. Let $Y_1(u)$ and $Y_2(u)$ be the number of individuals at risk in each group at time u;
$Y_j(u)= \sum_{K \in R_j} \mathrm{I}_{T \leq u}$, j=1,2 where R_j is the set of observation labels corresponding to group j. Let $\hat{\Lambda}_1(u)$ and $\hat{\Lambda}_2(u)$ be the Nelson Cumulative Hazard estimator for each group. The numerator of the log rank statistic can be expressed as a weighted difference between estimated hazard functions,

$$G = \int_0^\infty [w(u) \tfrac{Y_1(u)Y_2(u)}{Y_1(u)+Y_2(u)}](d\hat{\Lambda}_1(u)\text{-}d\hat{\Lambda}_2(u))$$

where $w(.) = 1$.

Stability of a statistic under censoring is an important requirement for splitting. If the statistic becomes more variable in censored data, it will tend to split in regions of heavier censoring. The performance of the Log-rank test statistic for partitioning censored data was investigated in a small simulation experiment; the results indicate that the Log-rank test performs well for splitting censored data. In addition, there are efficient updating algorithms for Log-rank test statistics (and permutation variances) for all possible split points. Updating algorithms are easily obtained, because the statistic can also be represented as the linear function :

$$G(s) = \textstyle\sum_i^n I_{X_{jt} \leq s}(\delta_i - \hat{\Lambda}_0(t_i))$$

where t_i is time under observation and X_{j_i} is the value for covariate j for individual i and s is the split point. This form of the statistic leads to an efficient updating formula for two split points s_1 and s_2 as,

$$G(s_1) - G(s_2) = \textstyle\sum_i^n I_{s_1 < X_{j_t} \leq s_2}(\delta_i - \hat{\Lambda}_0(t_i))$$

3.1 Splitting and Pruning for Censored Data

Recursive partitioning algorithms (RPART) can be described as follows. A rule splits the predictor space $\mathcal{X}$ into two disjoint regions. This rule is applied recursively to the data until the space has been split into many regions, each containing only a few observations. The partition can be represented as a binary tree T, where the set of terminal nodes $\tilde{T}$ corresponds to the partition of the co-variate space $\mathcal{X}$ into cardinality $|\tilde{T}|$ disjoint subsets. The components of the algorithm include rules for growing the tree, including the types of partitions that are permitted; rules for pruning the tree back; and rules for choosing a tree of the appropriate size.

3.1.1 Splitting Method

A split could be induced by any question of the form "Is $X \in S$, where $S \subseteq \mathcal{X}$"? [where $\mathcal{X}$ is the predictor space]

In CART several forms of splits are possible: splits of a single co-variate, splits on linear combinations of predictors, and Boolean combination splits. The simplest class of splits-splits on a single co-variate can be described by the following rules: (1) each split depends on the value of one predictor X_j; (2) if X_j is an ordered variable, then splits induced by questions of the form "Is $X_i \leq c$?" are allowed; and (3) if X_j is nominal with values in B = $b_1, b_2, ..., b_r$ then splits induced by any question of the form "Is $X_J \in S$?" where $S \subseteq B$ are allowed. Partitioning a node, h, involves finding the split, s, among all variables that maximizes some measure of improvement, G(s, h). In our case G(s, h) is usually a standardized two-sample log-rank test statistic. A tree is grown by finding the best split at each terminal node. The best split, s*, is the split such that,

$$G(s*, h) = max_{s \in s_h} G(s, h)$$

where s_h is the set of all possible splits of node h. If s* is not unique, then one of the maximal splits is arbitrarily chosen to partition the data. The same splitting rule is applied recursively to the resulting nodes until a large tree is grown with a small number of observations falling into each node. The tree is large and over fits the data at this point.Hence we describe an algorithm that efficiently finds optimally pruned sub-trees using a measure of the tree's performance based on the dissimilarity in survival between sibling nodes in the tree.

4 Theoretical Part and Splitting for conventional regression tree

In conventional tree method our data comes in a format where there are p co-variate and a response, for each of n observations, that is, (x_i, y_i) for i= 1,2,...n, with $x_i = (x_{i1}, x_{i2}, x_{i3}, ...x_{ip})$.The algorithm needs to decide the split point and the splits according to the data and also the shape of the tree.

Suppose we have a partition into M regions, say $R_1, R_2, ...R_M$, and we should model the response as a constant c_m in each region :

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

If we adopt as our criterion minimization of the sum of squares $\sum(y_i - f(x_i)^2)$,it is easy to see that the best $\hat{c_m}$ is just the average of y_i in region R_m :

$$\hat{c_m} = ave(y_i|x_i \in R_m)$$

Now finding the best binary partition in terms of minimum sum of squares is generally computationally infeasible.Hence we proceed with a greedy search method. Starting with all of our data, consider a splitting variable j and split point s, and define the pair of half planes

$$R_1(j, s) = [X|X_j \leq s] \text{ and } R_2(j, s) = [X|X_j > s]$$

Then we seek the splitting variable j and split point s that solve,

$$min_{j,s} \left[\min_{c_1} \sum_{xi \in R_1(j,s)} (yi - c_1)^2 + \min_{c_2} \sum_{xi \in R_2(j,s)} (yi - c_2)^2 \right]$$

For any choice of j and s, the inner minimization is solved by,

$$\hat{c_1} = avg(y_i|x_i \in R_1(j, s)) \text{ and } \hat{c_2} = avg(y_i|x_i \in R_2(j, s))$$

For each splitting variable, the determination of the split point s can be done very quickly and hence by scanning through all of the inputs, determination of the best pair (j,s) is feasible.

Having found the best split, we partition the data into the two resulting regions and repeat the splitting process on each of the two regions. Then this process is repeated on all of the resulting regions.

4.1 Choosing the Optimal Tree with Cost Complexity

Tree size is a tuning parameter governing the model's complexity, and the optimal tree size should be adaptively chosen from the data. One approach would be to split tree nodes only if the decrease in sum-of-squares due to the split exceeds some threshold. This strategy is too short-sighted, however, since a seemingly worthless split might lead to a very good split below it.

The strategy is to grow a large tree T_0 stopping the splitting process only when some minimum node size is reached.Then this large tree is pruned using cost complexity pruning.

4.1.1 Cost Complexity Pruning

We define a sub tree $T \subseteq T_0$ to be any tree that can be obtained by pruning T_0, that is, collapsing any number of its internal (non-terminal) nodes. We index terminal nodes by m, with node m representing region R_m.Letting,

$$N_m = \#[x_i \in R_m],$$

$$\hat{c_m} = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i)$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c_m})^2$$

we define the cost complexity criterion as ,

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

The main ideas is to find, for each α, the sub tree $T_\alpha \subseteq T_0$ to minimize $C_\alpha(T)$, The tuning parameter $\alpha > 0$ governs the trade-off between tree size and its goodness of fit to the data. Large values of α result in smaller trees T_α, and conversely for smaller values of α. As the notation suggests, with $\alpha = 0$ the solution is the full tree T_0. We discuss how to adaptively choose α below.

For each α one can show that there is a unique smallest sub tree T_α that minimizes $C_\alpha(T)$. To find T_α we use weakest link pruning,i.e, we successively collapse the internal node that produces the smallest

per-node increase in $\sum_m N_m Q_m(T)$, and continue until we produce the single-node(root) tree.

This gives a (finite) sequence of sub trees, and one can show this sequence must contain T_α.Estimation of α is achieved by 5 or 10 fold cross validation. We choose the value $\hat{\alpha}$ to minimize the cross-validated sum of squares.Our final tree is $T_{\hat{\alpha}}$.

5 Accuracy Metrics Calculation of Regression tree for Censored Data

We have used Concordance index and Integrated Brier Score metrics to evaluate the performance of the proposed prediction models and to compare them with existing ones. The Concordance Index (Harrell Jr et al., 1996) is widely used in survival analysis. It is similar to AUC in the sense that it measures the fraction of concordant or correctly ordered pairs of samples among all available pairs in the data set. The highest value of the metric is one (if the order is perfect), and the value of 0.5 means that the model produces completely random predictions. The following formula is used for calculating the concordance index:

$$CI = \frac{\sum_{i,j} 1_{T_j \leq T_i} 1_{\eta_j \leq \eta_i}}{\sum_{i,j} 1_{T_j \leq T_i}}$$

where T_j is the true time of the event, and η_j is the time predicted by the model. However, this metric is based only on the predicted time of the event, and it does not allow estimating the survival function. The value of CI does not change when the survival function is biased, although the predicted time is highly distorted compared to the true time.

To eliminate this problem, we use a metric called Integrated Brier Score (Murphy, 1973), (Brier and Allen, 1951), (Haider et al., 2020) based on the deviation of the predicted survival function from the true one (equal to 1 before the event occurs and 0 after that). The Brier Score (BS) metric (Brier and Allen, 1951) is used for estimating the performance of the prediction at a fixed time point t and is calculated in the following way:

$$BS(t) = \frac{1}{N} \sum_i (0 - S(t, x_i))^2 \text{ if } T_i \leq t$$

$$BS(t) = \frac{1}{N} \sum_i (1 - S(t, x_i))^2 \text{ if } T_i > t$$

where $S(t; x_i)$ is the prediction of the survival function at time t for observation x_i with event time T_i. To aggregate the BS estimates over all time moments, the Integrated Brier Score is used:

$$IBS = \frac{1}{t_{max}} \int_0^{t_{max}} BS(t) dt$$

6 Datasets

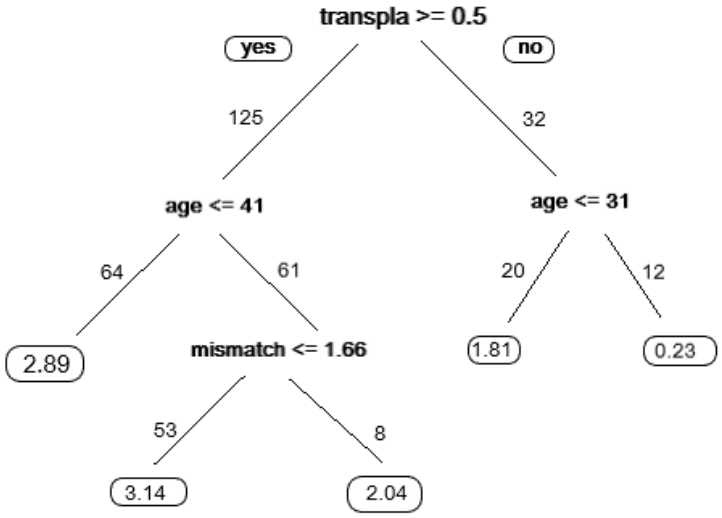
We have fitted the regression tree in some dataset from medical field where censoring is the main issue. All the data are collected from the book "SURVIVAL ANALYSIS Techniques for Censored and Truncated Data"(John P. Klein,Melvin L. Moeschberger).In each of the cases as we have attached tables consisting of various output figures as "n-split","relative error","x-error","x-std". Below we will discuss briefly the significance of these in the light of regression tree for censored data-

- n split- Number of splits in the tree. This column indicates how many times the tree was split during the training process.
- Rel Error- This is the error rate of the tree on the data. It's the proportion of misclassified or mispredicted observations in the data.
- X-error- Cross-validation error. This is an estimate of the error rate of the tree model on new, unseen data. It's calculated using cross-validation, where the dataset is divided into multiple subsets (folds), and the model is trained on some folds and tested on others. X-error (cross-validation error) is an important metric for assessing the predictive performance of survival tree model on new data. A lower cross-validation error indicates better predictive accuracy.
- X-std- Standard error of the cross-validation error. This column represents the standard deviation of the cross-validation error across different cross-validation folds. It gives a sense of the variability in the cross-validation estimates.

6.1 Stanford Heart Transplant Data

The response Y is survival time, where the survival time is the time (in days) until death due to rejection of the transplant heart. There are p=2 predictors, X_1, the age of the recipient, and X_2, a tissue mismatch score measuring recipient and donor tissue compatibility. One hundred fifty-seven cases were analysed, there being a 35% censoring rate. What has consistently emerged from the plethora of analysis is that age is the more significant predictor. The entire data is available under Survival library in R.

The proposed regression is attached below :



We had used the proposed Log-Rank Statistic Value which is implemented by the survival library function of R in this data and we got values according to the parameters such as age, mismatch score and transplant time. Result as follows:

Table 1: Log-Rank value

Log-Rank Value	nsplit	rel error	xerror	xstd
0.05453943	0	1.0000000	1.0081560	0.06939014
0.01465974	1	0.9454606	0.9912983	0.06954460
0.01233826	4	0.9014814	1.0903859	0.08673921
0.01000000	6	0.8768048	1.1021759	0.08984418

The tree chooses transplant as an important split here. From 157 patients 125 patients gone through the process and 32 withdrawn from the study. Those who had withdrawn from the study out of these patients only 20 patients were survived with 1.81% survival rate. And 12 whose age was greater than 31 only 12 patients survived with 0.23% survival rate.

at the left hand side we are getting patients who had undergone through the process of transplantation. Out of 125 patients 64 patients who were below 41 years age survived with a higher probability of 2.89%. And Those who are above 41 years there is a problem of tissue mismatch factor with the donor. But 53 were survived with a higher chance of survival rate whose tissue mismatch score was less than 1.66. And 2.04% was the survival rate of the rest 8 patients whose tissue mismatch score was above 1.66.

6.2 Death Times of Psychiatric Patients

Woolson (1981) has reported survival data on 26 psychiatric inpatients admitted to the University of Iowa hospitals during the years 1935–1948. This sample is part of a larger study of psychiatric inpatients discussed by Tsuang and Woolson (1977). Data for each patient consists of age at first admission to the hospital, sex, number of years of follow-up (years from admission to death or censoring) and patient status at the follow up time.

Gender	*Age at Admission*	*Time of Follow-up*
Female	51	1
Female	58	1
Female	55	2
Female	28	22
Male	21	30$^+$
Male	19	28
Female	25	32
Female	48	11
Female	47	14
Female	25	36$^+$
Female	31	31$^+$
Male	24	33$^+$
Male	25	33$^+$
Female	30	37$^+$
Female	33	35$^+$
Male	36	25
Male	30	31$^+$
Male	41	22
Female	43	26
Female	45	24
Female	35	35$^+$
Male	29	34$^+$
Male	35	30$^+$
Male	32	35
Female	36	40
Male	32	39$^+$

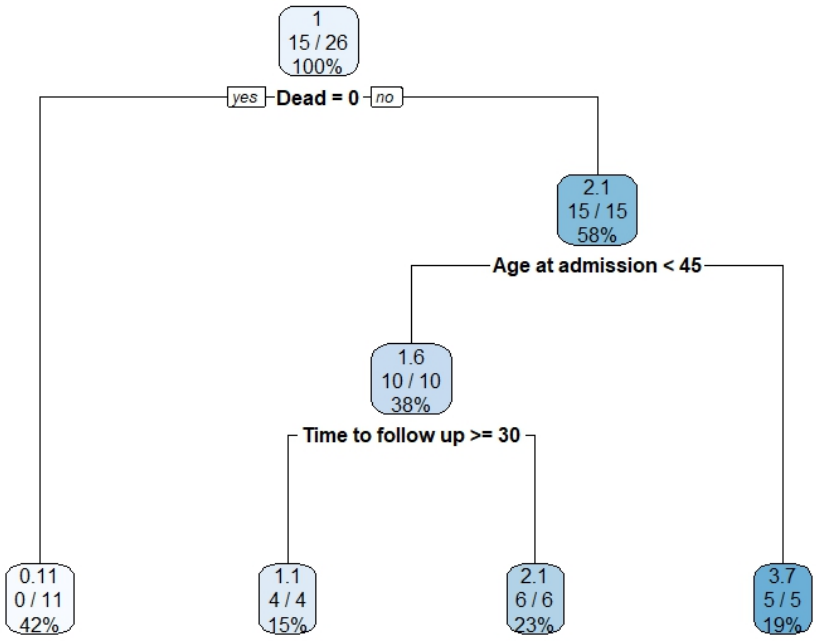
$^+$Censored observation

' According to the data we tried to fit the regression tree here with Log-Rank split as main criterion and the table attached below :

Log-rank value	nsplit	rel error	xerror	xstd
0.67885671	0	1.0000000	1.0897031	0.14526219
0.11627436	1	0.3211433	0.3778317	0.09881194
0.04051891	2	0.2048689	0.2858834	0.06559389
0.01000000	3	0.1643500	0.2474773	0.06408628

Table 2: Log-Rank Table for Death Times of Psychiatric Patients

Proposed regression tree for the above data is attached below:



In this case at first according to the output the main split came out as Death of a patient.Out of 26 patients were found dead that is 42% observation supposed to be dead. Then on the right side of the tree we are looking at alive patients and patients with age below 45 or above 45. Then further we splitted alive patients who took admission in the study before 45 years into time to follow up group. That is patients who followed the study for more than 30 months or less than 30 months.In the final nodes we got the predicted percent-

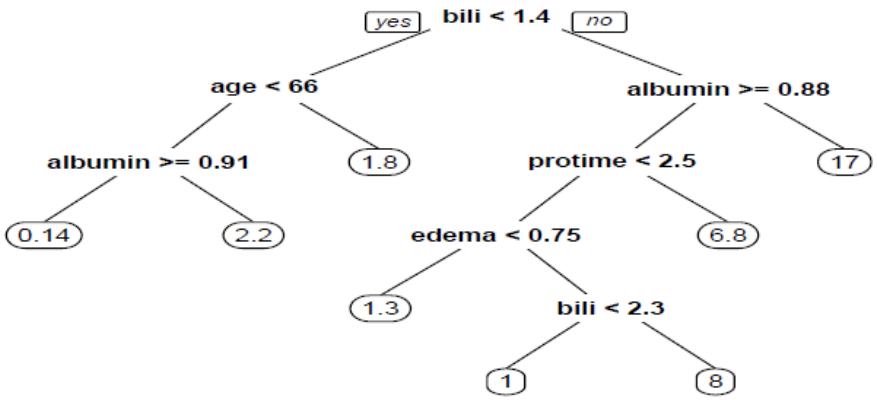
age of patients who survived with covariates age and time to follow up of the study.

6.3 The Mayo Clinic Primary Biliary Cirrhosis Data

The pbcseq dataset in the survival package is used as an example to illustrate fitting survival trees with time-varying covariates. These data were obtained from 312 patients with primary biliary cirrhosis (PBC) enrolled in a double-blind, placebo-controlled, randomized trial conducted between January, 1974 and May, 1984 at the Mayo Clinic to evaluate the use of D-penicillamine for treating PBC. A comprehensive clinical and laboratory database was established on each patient. Follow-up was extended to April, 1988, by which time 140 of the patients had died and 29 had undergone orthotopic liver transplantation. These patients generated 1,945 patient visits that enable us to study the change in the prognostic variables of PBC. Several variables of this data are as follows :

- age: in years
- albumin: logarithm of serum albumin (g/dl)
- alk.phos: alkaline phosphates (U/liter)
- ascites: presence of ascites
- ast: aspartate aminotransferase(U/ml)
- bili: logarithm of serum bilirubin (mg/dl)
- chol: serum cholesterol (mg/dl)
- edema: 0-no edema, 0:5-untreated or successfully treated, 1-edema despite diuretic therapy
- hepato: presence of hepatomegaly or enlarged liver
- platelet: platelet count
- protime: logarithm of prothrombin time, standardized blood clotting time
- spiders: presence or absence of spiders

The proposed regression tree is attached below for this data:



It is clear that the tree chooses bili, age, albumin and protime as important risk factors and split. In the PBC data case, it is safe to say that age, protime, albumin are predictive risk factors for survival time of individuals with primary biliary cirrhosis, since they are identified by all of the models, while the importance of other potential risk factors such as edema may be decided by further analysis. As mentioned above in the terminal nodes we are getting the final prediction of the survival rate of individuals with primary biliary cirrhosis. From above we have splitted the tree with sirum bilirubin index (billi) which is less than 1.4 mg/dl or not ? Then age came out as an important split. We are looking at patients whose age is less than 66 or not. In the terminal node it has been observed that only 1.8% has been survived who are above 66 years. Then patients whose age is less than 66 we will observe the albumin portion. In extreme left node we have observed patients with albumin greater than 0.91 g/dl survived with a very lower survival rate where as patients with low albumin and age below 66 survived with a little bit more probable ratio. In this manner in the right side of the tree protime and edema are another two important covariates along with age, albumin and bili.

7 Accuracy Metric for Regression Trees

As discussed earlier that as a measure of accuracy we will look into the concordance index of the regression tree for censored data and as a better accuracy measure the value of Brier score is calculated for each individual tree for three sets of data.

Dataset	Concordance Index	Brier Score
Stanford Heart Transplant data	0.64533	**0.16498**
Death Time for Psychiatric Patients	0.58500	**0.19119**
PBC Data	0.65284	**0.21341**

Table 3: Regression Tree Performance

In each and every case from the aforementioned table we are observing that brier score outperformed Concordance index in every case.

8 Bibliography

- Fu, W. and Simonoff, J.S. (2017a). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. Biostatistics 18 (2), 352-369
- Fu, W. and Simonoff, J.S. (2017b). Survival trees for Interval Censored Survival data. Statistics in Medicine 36 (30), 4831-4842
- Therneau, T., Atkinson, B., Ripley, B. and Ripley, M.B., (2015). rpart: Recursive Partitioning and Regression Trees.
- Hothorn, T. and Zeileis, A., (2015). partykit: A Modular Toolkit for Recursive Partytioning in R.
- herneau, T., (2015). survival: A Package for Survival Analysis in S. version 2.38.
- Milborrow, S., (2011). rpart.plot: Plot 'rpart' models: An Enhanced Version of 'plot.rpart'.
- Survival Trees by Goodness of Split,Michael LeBlanc and John Crowley,Journal of the American Statistical Association , Jun., 1993, Vol. 88, No. 422 (Jun., 1993), pp. 457-467
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), Classification and Regression Trees, Belmont, CA: Wadsworth International Group.
- Aalen, O. (1978). Nonparametric inference for a family of counting processes. The Annals of Statistics, 6:701726.
- Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18:2529 2545.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,
- Hastie, T., Tibshirani, R., Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, Springer.

- Antionio Ciampi, Abdissa Negassa, and Zihyi Lou. Tree-structured prediction for censored survival data and the cox model. Journal of Clinical Epidemiology, 48(5):675689, 1995.
- L.A. Escobar and Jr. Meeker W.Q. Assessing influence in regression analysis with censored data. Biometrics, 48:50728, 1992.
- Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, University of California, Berkeley, 1981.